

THE EFFECTS OF MEASUREMENT ERRORS ON RELATIVE RISK REGRESSIONS

BEN G. ARMSTRONG

Armstrong, B. G. (School of Occupational Health, McGill U., Montreal, Quebec, Canada H3A 1A3). The effects of measurement errors on relative risk regressions. *Am J Epidemiol* 1990;132:1176-84.

This paper concerns the effects of random error in numerical measurements of risk factors (covariates) in relative risk regressions. When not dependent on outcome (nondifferential), such error usually attenuates relative risk estimates (shifts them toward one) and leads to spuriously narrow confidence intervals. The presence of measurement error also reduces precision of estimates and power of significance tests. However, significance levels obtained by using the approximate measurements are usually valid and as powerful as possible given the measurement error. The attenuation in risk estimate depends not only on the size (variance) of the measurement error, but also on its distributional form, on whether it is dependent on the true level of the risk factor (whether it is of "Berkson" type), on the variance and distributional form of true levels of the risk factor, on the functional form of the regression (exponential or linear), and on the confounding variables included in the model. Error in measuring confounding variables leads to loss of control of confounding, leaving residual bias. Uncomplicated techniques of correcting the effects of measurement error in simple models in which distributions are assumed normal are available in the statistical literature. For these corrections, information on measurement error variance is required. Some approaches appropriate for more general models have been proposed, but these appear to be insufficiently developed for routine application.

regression analysis; risk; statistics

Once a causal association between a risk factor (covariate) and disease occurrence is considered likely, interest often focuses on the quantitative relation with risk—for a given change in the level of the covariate, what is the change in risk? For occupational and environmental pollu-

tants, this exposure-response relation is a necessary basis from which standards of exposure carrying "acceptable" risks can be determined.

All estimates of relations between covariates and disease made from epidemiologic studies require estimates of the covariate for persons or groups of persons in the study. There is almost always a possibility of error in these measurements. We are concerned here with covariate measurements on a numerical scale. In this case, the amount by which the measurement differs from the "true" level (often hypothetical) is called measurement error. The notion of truth here is not absolute (such as a biologically effective dose), but will depend upon context. Usually it will represent

Received for publication March 31, 1989, and in final form June 5, 1990.

From the School of Occupational Health, McGill University, Montreal, Quebec, Canada.

Reprint requests to Dr Ben G Armstrong, School of Occupational Health, McGill University, 1130 Pine Avenue West, Montreal, Quebec, Canada H3A 1A3.

This research was supported by the Canadian National Health Research and Development Program and National Sciences and Engineering Research Council, and by the Institut de Recherche en Santé et en Sécurité du Travail du Québec

the level that would have been observed had the subject worn an accurate personal dosimeter throughout his or her life. The error may be systematic (constant for all persons in the study), random (so that the average of many replicated measurements converges to the true covariate value), or a combination of both. This paper concerns random error.

Willett (1) argues the importance of random covariate measurement error in epidemiology. The aim of this paper is to provide an accessible overview of the extant statistical results on the effects of this error, and to discuss their relevance for typical epidemiologic contexts.

A SIMPLE MEASUREMENT ERROR MODEL FOR RELATIVE RISK REGRESSION

For cohort studies with times of disease occurrence observed, disease rate (hazard) ratios may be regressed on covariates using Cox regression. For case-control studies in which controls are chosen from among persons surviving without disease to the age at which the case is diagnosed (i.e., individually matched or closely stratified by age), essentially the same model is appropriate. To be consistent with the relevant statistical literature (2), we call these "relative risk regression models," although "risk" here does not imply cumulative incidence, as it does in some epidemiologic literature, but rather incidence density.

In the relative risk regression model, disease risk (incidence density) in persons with covariate value x is specified relative to risk in persons with the same age (t) but with the covariate at zero ($x = 0$). This relative risk (RR) is independent of age and depends on x through a regression coefficient β :

$$\text{RR}(x, t) = r(\beta, x). \quad (1)$$

Although the variable t is usually taken as age, other choices (e.g., time since surgery) are possible. The covariate x may be a summary history of levels of the risk factor experienced up to age t , and thus change with t (e.g., cumulative exposure).

The function $r(\cdot)$ can take several forms; the most popular is the exponential (closely related to the logistic model):

$$\text{RR}(x, t) = \exp(\beta x). \quad (2)$$

The exponential model is assumed in this paper unless specified otherwise. In this model, $\exp(\beta)$ represents the proportional change in risk per unit change in x , which is the same whatever the original value of x . In this sense, this model is not affected by choice of origin (zero) for x .

The effects of measurement error

Defining x as the true covariate and z as the (approximate) measurement of x , we may write:

$$z = x + e, \quad (3)$$

where e is an additive measurement error. In the usual ("classical") measurement error model, we assume in addition to error being nondifferential and wholly random (nonsystematic) that it has a distribution independent of the true covariate x . To obtain simple results, we further assume that the distribution of error e is normal (Gaussian) with variance σ_e^2 and that the age-specific distribution of true covariate x is normal with mean μ_x (which may depend on age) and variance σ_x^2 which is independent of age.

Under these assumptions, the relation between relative risk and approximate z retains the same exponential form as that between relative risk and true x , but the "naive" regression parameter, β^* , is attenuated (shifted toward zero), i.e., $\text{RR}(\beta^*, z) = \exp(\beta^* z)$, with $|\beta^*| < |\beta|$. Thus, the proportional change in risk per unit change in observed approximate level of the covariate is less than that per unit change in the true level.

The naive parameter β^* is connected to β by the identity:

$$\beta^* = R\beta, \quad (4)$$

where $R = \sigma_x^2 / (\sigma_e^2 + \sigma_x^2)$. If the variance of observed measurements ($\sigma_x^2 + \sigma_e^2$) is written σ_z^2 , R may be reexpressed as σ_x^2 / σ_z^2 or

$(\sigma_z^2 - \sigma_e^2)/\sigma_z^2$. R is known as the reliability of z as a measure of x and is equal to the square of the correlation of x and z in the study population. The attenuation is precisely that for the slope in a simple linear regression for a numerical outcome when the regressor is subject to measurement error (3). The monograph by Fuller (4) contains an authoritative discussion of this and many related models for numerical outcomes.

Table 1 shows the attenuation R in the relative risk regression parameter β due to measurement error with standard deviation (σ_e) up to two times that of the true covariate (σ_x). (It is this relative variation (σ_e/σ_x) which determines R , and hence attenuation.) For many contexts, the results are reassuring: Measurement error with a standard deviation as large as half of the standard deviation of the true x results in an observed regression parameter which is on average 80 percent of the true value—a minor bias by most epidemiologic standards (1).

The naive significance tests against $\beta^* = 0$ (e.g., score, Wald, or likelihood ratio chi-square tests) using the crude data and ignoring measurement error are valid and efficient tests for $\beta = 0$ (5). The tests will, however, have less power than those based on the true covariate, since their efficiency is equal to the reliability R of the measured covariate (6).

The above and further related results are given in recent publications in the statistical literature (7–11).

Parallels with misclassification

The results given above are consistent with well-known results for categorical (in particular dichotomous) covariate mea-

sures subject to misclassification that is nondifferential. Relative risks (odds ratios) are shifted toward 1; tests for association remain valid, but with reduced power compared with those using data without misclassification. Chen (12) provides a recent review from a statistical perspective. For ordered categories to which numerical scores are assigned and over which trends in risk are to be estimated, the measurement error models discussed above may be applied, since the results are essentially the same as those for continuous covariates.

MODELS WITH MANY COVARIATES:
ERRORS IN CONFOUNDING VARIABLES

Even when we are mainly interested in a single risk factor, we often wish to control in the analysis for confounding by one or more additional factors. Further, we may be interested in investigating several factors simultaneously. Much of the popularity of relative risk regression models lies in their capacity to carry out these functions.

Measurement error may be present in the observed factors of interest and in confounders. In an extension of the above model, we assume a multivariate normal distribution for true covariates and (independently) for errors. This model implies that measurement error for each covariate is independent of *all* true covariates, so that, for example, error in the factor of interest does not depend on the true level of a confounder. Under this assumption, results for the exponential model analogous to expression 4 are available (11). Below we consider the qualitative implication of these results. For simplicity, we assume that there is just one confounder, although the results given extend qualitatively to many confounders.

If only the factor of interest is subject to

TABLE 1
Attenuation in estimated relative risk parameter due to normal measurement error

Error σ_e/σ_x^*	0.00	0.10	0.20	0.30	0.40	0.50	0.75	1.00	1.50	2.00
Attenuation (R)	1.00	0.99	0.96	0.92	0.86	0.80	0.64	0.50	0.31	0.20

* Ratio of standard deviation of errors e to standard deviation of the true covariate x .

error, the regression parameter for this covariate is attenuated, as in the univariate case, but this attenuation is more pronounced. Using subscripts 1 and 2 to denote variables and parameters associated with the factor of interest and the confounder, respectively, the naive coefficient of interest $\beta_1^* = R_{1|2}\beta_1$, where $R_{1|2}$ is obtained by replacing $\sigma_{z_1}^2$ in the formula $R = (\sigma_{z_1}^2 - \sigma_{z_1}^2)/\sigma_{z_1}^2$ by $\sigma_{z_1|z_2}^2$, the variance of z_1 conditional on the confounder z_2 (see Appendix for formulae and proof). This variance can be estimated from the variance of the residuals if z_1 is regressed on z_2 .

The increase in attenuation occurs even if the second factor is not associated with the disease and has thus been unnecessarily included as a confounder. This may happen, for example, in occupational studies when length of service is unnecessarily included as a possible confounder (to guard against "survivor effect" confounding) and interest is in a measure, for example, cumulative exposure, which is strongly correlated with length of service. Of course, if length of service is a risk factor independent of the effect of cumulative exposure, not including it will introduce confounding bias.

If the confounder is measured with error, then its inclusion in the regression can only partially control for its effect, leaving "residual" confounding. Thus, if the variable of interest is measured without error, the naive coefficient of interest β_1^* will lie between β_1 and the confounded coefficient β_1^\dagger which would be obtained if the confounder were omitted. With notation as before, $\beta_1^* = R_{2|1}\beta_1 + (1 - R_{2|1})\beta_1^\dagger$ (see Appendix). Tests of the statistical significance of the covariate of interest which do not account for this error are not valid. Alternative tests are available (11). If the variable of interest is also subject to error, then the residual confounding is coupled with the pronounced attenuation discussed above: $\beta_1^* = R_{2|1}(R_{1|2}\beta_1) + (1 - R_{2|1})\beta_1^\dagger$ (see Appendix). If the two errors are correlated (as is likely, for example, in estimating dietary intakes of components present in

the same foodstuffs), the extent and direction of the changes may be altered.

The effects of misclassification in dichotomous measures of a variable of interest and a confounder (13) and those of measurement error on partial correlation coefficients (14) are similar qualitatively to those described above.

CORRECTING FOR ATTENUATION DUE TO MEASUREMENT ERROR

Method

If R (or σ_e^2 , from which R may be deduced) is known, a maximum likelihood estimate and confidence interval for the "true" β in the univariate exponential relative risk model, expression 2, may be obtained by applying a correction factor ($1/R$) to the naive maximum likelihood estimate $\hat{\beta}^*$, and its confidence interval may be obtained from the observed data z , i.e.,

$$\hat{\beta} = \hat{\beta}^*/R. \quad (5)$$

$\hat{\beta}^*$ will usually be obtained using a conditional logistic or Cox regression package. There is an equivalent matrix formula for correcting for measurement error in the multivariate model (11).

Usually R must be *estimated*. We may do this by estimating σ_e^2 from a validity study in which approximate measures of the covariate are compared with a "gold standard," or from a reliability study of repeated independent approximate measures of the covariate in the same individuals. The required independence of repeated measurements (absence of error systematic to an individual) is often difficult to achieve (1). Given σ_e^2 , we may estimate R as $(\sigma_z^2 - \sigma_e^2)/\sigma_z^2$, where σ_z^2 is the (age-specific) variance of z in the *main study*.

Where information from which σ_e^2 is estimated is sparse, so that uncertainties in σ_e^2 and hence R may not be ignored, the corrected confidence limits should be refined so that the extra uncertainty is reflected in a wider interval (11). In addition, the "corrected" estimate obtained by

expression 5 may retain important bias because of the small size of the validity or reliability study. This and the additional imprecision introduced into the estimates may outweigh the probable reduction in attenuation. Because of these difficulties, the best approach may often be to carry out sensitivity analyses in which the consequences of measurement errors of various magnitudes are investigated, using the method in which R is assumed known. Choices of R used in these analyses may be guided by general plausibility considerations, as well as any data available.

Example

As part of a broader study, dietary intakes were estimated, using dietary history questionnaires, in 171 cases of colon cancer and 171 controls, individually matched for age and neighborhood (15-17). The naive exponential relative risk model was fitted for this illustration to daily fat intake (in grams). The first row of table 2 shows the naive estimate $\hat{\beta}^*$, obtained from a conditional logistic regression package. Subsequent rows show the effects of correcting this for measurement error by using expression 5, assuming the normal classical model. Corrections for mild ($\sigma_e/\sigma_x = 0.1$) to quite severe ($\sigma_e/\sigma_x = 1.0$) measurement error are shown. Table 2 also shows the relative risk predicted because of a difference in true fat intake of 50 g (exp (50β)). As stated above, the significance level ($p = 0.04$) is unaltered by correction for measurement error. This also explains why the lower 95 percent confidence interval changes very little on correction.

A companion validity study compared dietary histories of 16 volunteers with detailed weighed food records kept by their spouses (18). If the latter is taken as the gold standard, measurement error variance σ_e^2 is estimated as 1,284 g² (95 percent confidence interval 710-2,970). The stratum-specific variance of observed fat intake σ_x^2 was calculated as half the variance of the 171 differences between intakes in cases

TABLE 2
Colon cancer and fat intake estimates of relative risk regression parameters corrected for normal measurement error

Error* $\frac{\sigma_e}{\sigma_x}$	Corrected estimate (95% confidence interval)†	
	$\beta \times 10^2$	Relative risk for 50 g difference
0.00†	4.6 (0.1-9.1)	1.3 (1.0-1.6)
0.10	4.6 (0.1-9.2)	1.3 (1.0-1.6)
0.20	4.8 (0.1-9.5)	1.3 (1.0-1.6)
0.30	5.0 (0.1-9.9)	1.3 (1.0-1.6)
0.40	5.3 (0.1-10.6)	1.3 (1.0-1.7)
0.50	5.8 (0.1-11.4)	1.3 (1.0-1.8)
0.60	6.3 (0.1-12.4)	1.4 (1.0-1.9)
0.70	6.9 (0.1-13.6)	1.4 (1.0-2.0)
0.80	7.5 (0.2-14.9)	1.5 (1.0-2.1)
0.90	8.3 (0.2-16.5)	1.5 (1.0-2.3)
1.00	9.2 (0.2-18.2)	1.6 (1.0-2.5)

* Ratio of standard deviation of errors e to standard deviation of true covariate x .

† The first row gives uncorrected (naive) estimates

and matched controls, 2,667 g². We may thus estimate σ_x^2 as 2,667 - 1,284 = 1,383 g², σ_e/σ_x as $\sqrt{(1,284/1,383)} = 0.96$, R as 1,383/2,667 = 0.52, and β as 4.6/0.52 = 8.8. However, since the sampling uncertainty in σ_e^2 is substantial, R cannot be considered to be known to be this value. This estimate of β is thus not maximum likelihood, and confidence intervals calculated as above would not reflect uncertainty in σ_e^2 . Further, note that any confounding or other bias present in the naive estimate $\hat{\beta}^*$ will be magnified by this "correction."

To keep the illustration simple, we have not considered additional possibly confounding variables. Armstrong et al. (11) give a fuller analysis and discussion of this example, including confounders in particular.

OTHER MEASUREMENT ERROR MODELS

Unfortunately, the simple model for measurement error described above rarely applies exactly to data arising in epidemiologic studies. Often the results from this model will be sufficiently valid to examine the approximate impact of measurement error even when assumptions are not met

exactly, but if precise corrections are required or if departures from model assumptions are substantial, other models should be considered.

Linear relative risk models

Investigators may prefer alternatives to the exponential form (expression 2) of the relative risk model. In particular, the linear model, $RR = 1 + \beta x$, is consistent with some suggested models for carcinogenesis and provides a better fit to data from some studies of environmental exposures and cancer (19). It is less attractive to assume a normal distribution for errors and covariates in the linear relative risk model, in which choice of zero on the x scale affects the model and which is only sensibly applied to covariates taking only nonnegative values. However, if we proceed with this working assumption, we can obtain simple approximate results. We require the further assumption that mean true covariate $\mu_x(t)$ is independent of age t . Then the naive relation remains linear, with coefficient $\beta^* = [R/\{1 + \beta\mu_x(1 - R)\}]\beta$ (7-9). Because the normal assumption is usually unrealistic, this relation should be used with special caution. Nonnormal error models, discussed below, are more useful for the linear model.

Nonnormal distribution of measurement error and of covariates

The results so far depend on measurement error and true covariates being distributed normally. This is clearly often an imperfect model since covariates are often measured on a scale with an origin at zero. Further, the distributions of many covariates (for example, cumulative exposures to airborne contaminants) are skewed to the right, often approximating the lognormal distribution. In addition, measurement errors may be better represented in a multiplicative rather than an additive model: $z = xe$, with $E(e) = 1$. One such model has measurement error and the true covariate following a lognormal distribution, inviting the representation: $\log(z) = \log(x) + \log(e)$.

In this model, taking logarithms has the effect of returning to the additive formulation, with normal distribution for measurement errors and the true covariate. If $\log(x)$ is taken as the variable of interest, the simple methods of the preceding section may thus be applied. In the exponential model, this implies a true relation: $RR(x, t) = \exp(\beta \log(x)) = x^\beta$, with measurement error attenuating β . Since β appears as a power of x , measurement error will alter the shape as well as the magnitude of the slope of relations of relative risk to the covariate (7, 20). For example, a quadratic relation of lung cancer risk with true pack-years of exposure to tobacco smoke could be distorted to a linear form due to this type of measurement error (21).

The form of the relation between risk and the (original true) covariate implied by this model may not fit the data or prior assumptions of mechanisms, however. Models in which covariates or measurement errors are skewed but relative risk is related to the covariate on its original scale have been discussed explicitly only briefly in the literature (8), although other more general approaches (7, 9-11) may be adaptable to this situation. Further discussion is beyond the scope of this paper.

Under nonnormal measurement error models, naive tests of $\beta^* = 0$ using the crude data remain valid tests of $\beta = 0$, but are no longer as powerful as some alternatives (5).

Covariate distribution dependent on age

The validity of expressions 4 and 5 depends on the age-specific distributions of true covariates being normal and the variance σ_x^2 being independent of age. However, if the covariate affects the disease of interest (i.e., if $\beta \neq 0$) or mortality from other causes, disease or early deaths among highly exposed subjects will cause a dependence of the distribution of the true covariate on age (7, 9). In many cases, the dependence is weak, so that the simple expressions 4 and 5 remain good approximations.

Predictive and structural relations

We have assumed in the above discussion that the parameter of interest is β , the true regression slope in the population, rather than β^* , the naive regression slope in the population. However, this may not always be true. For prediction of relative risk in a person drawn from a population *with the same distributions of the true covariate and measurement error as the study group*, the naive β^* should be used (22). For this reason, the regression on z involving β^* is called the "predictive" relation in contrast to that on x involving β , which is known as the "structural," or in a slightly different context the "functional" relation (23). To predict risk in contexts in which either of these distributions is changed, we can obtain a new predictive coefficient β^* from β if the new σ_e^2 and σ_x^2 are known.

Berkson models

Expression 3, in which error e is independent of x , is generally called the classical measurement error model. There is an alternative which was originally proposed for experimental situations in which the experimenter attempts to set a covariate at a target value z , but because of imprecise control its true value x may be higher or lower than z . If the experiment was replicated many times with the same target z , the true covariate could often be expected to be distributed with mean z . This situation may be represented $x = z + e$, with e independent of z , and $E(x|z) = z$. This is called the Berkson model, after Berkson (24), or sometimes the "control knob" model.

Remarkably, in the linear relative risk regression model with normal measurement error of the Berkson type, the naive parameter β^* is equal to the true β —there is no attenuation. However, $\hat{\beta}^*$ estimated from the approximate $\{z\}$ has a higher standard error than $\hat{\beta}$ estimated from the true $\{x\}$, and power to test $\beta = 0$ is reduced. In the exponential model with normal Berkson error, there is also no attenuation providing that the variance σ_e^2 of e does not

depend on z (7, 9). To the extent that such a dependency exists, the naive and true regression parameters will differ in the exponential model, not necessarily in the direction of attenuation. Prentice (7) discusses a model for error of this kind in which $|\beta^*| > |\beta|$.

The relevance of the Berkson model for epidemiology is in its application to certain observational as well as experimental studies. For example, a single estimate z of exposure is often taken to apply to all workers with the same job title. For each worker, this is an approximation of their true exposure x . If z is the mean true exposure ($z = \bar{x}$) for this job title, the situation is analogous to the control knob context above: e (equal to $x - \bar{x}$) is independent of z , and Berkson's model applies. However, the single estimate z of exposure is itself usually an approximation of the true mean exposure \bar{x} among workers with the job title. The difference ($\bar{x} - z$) between this true mean and the exposure estimate will often conform to a classical measurement error model. Thus, most relative risk regressions drawing on this type of exposure estimation will be subject to both Berkson and classical types of measurement error.

A similar type of error structure pertains if numerical measures of a covariate are used to divide subjects into categories, and mean level in each category is used in relative risk regressions. The approximation involved in this procedure introduces additional error of the Berkson type, but leaves the classical error structure essentially the same. Further, if the width of bins neither increases nor decreases with z , σ_e^2 is unlikely to depend on z , so that in this case the exponential as well as the linear model grouping does not cause attenuation. Grouping does reduce precision, however.

Logistic and other binary regression models

When "logistic" regression is used to analyze case-control data stratified by age, it is the (exponential) relative risk model which is assumed in the underlying popu-

lation (2), so that the above results remain applicable. When logistic or other binary regression methods are applied to prevalence or simple cohort studies with time to disease occurrence not observed, the relative risk model described above no longer applies. More precise, but more computationally arduous methods of accounting for measurement error effects in these type of binary regression models are discussed by several authors (25–33). When proportions with the disease are small, however, the distinction between these and relative risk models is minor (2), so that the effects of measurement error would be expected to be close to those described above.

REFERENCES

1. Willett W. An epidemiologic perspective on exposure measurement error. *Stat Med* 1989;8:1031–65.
2. Prentice RL, Farewell VT. Relative risk and odds ratio regression. *Ann Rev Public Health* 1986; 7:35–58.
3. Snedecor GW, Cochran WG. *Statistical methods*. Ames, Iowa: Iowa State University Press, 1967: 164–7.
4. Fuller WA. *Measurement error models*. New York: John Wiley & Sons, 1987.
5. Tosteson TD, Tsiatis AA. The asymptotic relative efficiency of score tests in the generalized linear model with surrogate covariates. *Biometrika* 1988; 75:507–14.
6. Lagakos S. Effects of misspecification and mismeasuring explanatory variables on tests of their association with a response variable. *Stat Med* 1987;7:257–74.
7. Prentice RL. Covariate measurement error and parameter estimation in a failure time regression model. *Biometrika* 1982;69:331–42.
8. Armstrong BG, Oakes D. The effects of approximation in exposure assessments on estimates of exposure response relationships. *Scand J Work Environ Health* 1982;8 (suppl. 1):20–3.
9. Clayton DG. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: Dwyer JH, Lippert P, Feinleib M, et al., eds. *Statistical models for longitudinal studies of health*. New York: Oxford University Press, 1988.
10. Pepe M, Self SG, Prentice RL. Further results on covariate measurement errors in cohort studies with time to response data. *Stat Med* 1989;8:1167–78.
11. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Stat Med* 1989;8:1151–65.
12. Chen TT. A review of methods for misclassified categorical data in epidemiology. *Stat Med* 1989; 8:1095–106.
13. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; 112:564–9.
14. Kupper LL. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol* 1984;120:643–8.
15. Jain M, Cook GM, Davis FG, et al. A case-control study of diet and colo-rectal cancer. *Int J Cancer* 1980;32:757–68.
16. Miller AB, Howe GR, Jain M, et al. Food items and food groups as risk factors in a case-control study of diet and colorectal cancer. *Int J Cancer* 1983;32:155–61.
17. Howe GR, Miller AB, Jain M. Re: "Total energy intake: implications for epidemiologic analysis." (Letter). *Am J Epidemiol* 1986;124:157–9.
18. Jain M, Howe GR, Johnson KC, et al. Evaluation of a diet history questionnaire for epidemiologic studies. *Am J Epidemiol* 1980;111:212–19.
19. Thomas DC. General relative risk models for survival time and matched case-control analysis. *Biometrics* 1981;37:673–86.
20. Prentice RL. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J Am Stat Assoc* 1986;81:321–7.
21. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular and lifelong nonsmokers. *J Epidemiol Community Health* 1986;32:303–13.
22. Madansky A. Fitting straight lines when both lines are subject to error. *J Am Stat Assoc* 1959; 54:173–204.
23. Kendall M, Stuart A. *The advanced theory of statistics*. Vol 2. New York: MacMillan, Inc., 1979:399–443.
24. Berkson J. Are there two regressions? *J Am Stat Assoc* 1950;45:164–80.
25. Carroll RJ, Spiegelman CH, Lan KK, et al. On errors-in-variables for binary regression models. *Biometrika* 1984;71:19–25.
26. Stefanski LA, Carroll RJ. Covariate measurement error in logistic regression. *Ann Stat* 1985;13: 1335–51.
27. Armstrong BG. Measurement error in the generalised linear model. *Comm Stat Simulation Comput* 1985;14:529–44.
28. Schafer DW. Covariate measurement error in generalised linear models. *Biometrika* 1987;74:385–91.
29. Whittemore AS, Grosser S. Regression models for data with incomplete covariates. In: Moolgavkar SH, Prentice RR, eds. *Modern statistical methods in chronic disease epidemiology*. New York: John Wiley & Sons, 1986:19–34.
30. Burr D. On error-in-variables in binary regression—Berkson case. *J Am Stat Assoc* 1988;83: 739–43.
31. Whittemore AS, Keller JB. Approximations for regression with covariate measurement error. *J Am Stat Assoc* 1988;83:1057–66.
32. Tosteson TD, Stefanski LA, Schafer DW. A measurement error model for binary and ordinal regression. *Stat Med* 1989;8:1139–48.
33. Rosner B, Willett W, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person

- measurement error Stat Med 1989;8:1051-70
- 34 Snedecor GW, Cochran WG. Statistical methods. 6th ed. Ames, IA: Iowa State University Press, 1967:395.

APPENDIX

Attenuation of the regression parameter in the presence of a confounder

In addition to the notation introduced in the text, we denote the covariance of z_1 and z_2 , as σ_{z_1,z_2} , and the conditional variance of z_1 on z_2 as $\sigma_{z_1|z_2}^2$, so that (given bivariate normality) $\sigma_{z_1|z_2}^2 = \sigma_{z_1}^2 - \sigma_{z_1,z_2}^2/\sigma_{z_2}^2$. Other covariances and conditional variances are denoted likewise. Then the conditional reliability $R_{1|2} = \sigma_{z_1|z_2}^2/\sigma_{z_1}^2$ and $R_{2|1} = \sigma_{z_2|z_1}^2/\sigma_{z_2}^2$. If we further note that our measurement error model implies $\sigma_{z_1,z_2}^2 = \sigma_{x_1,x_2}^2$ and substitute zero covariance between the two measurement errors into expression 4 of reference 11, we obtain.

$$\beta_1^* = R_{1|2}\beta_1 + (1 - R_{2|1})(\sigma_{x_1,x_2}/\sigma_{x_1}^2)\beta_2 \quad (6)$$

The value of β_1^* obtained by omitting the confounder entirely from the regression can be obtained by letting $R_{2|1} \rightarrow 0$, giving

$$\beta_1^\dagger = R_{1|2}\beta_1 + (\sigma_{x_1,x_2}/\sigma_{x_1}^2)\beta_2 \quad (7)$$

Substituting back into expression 6 gives

$$\beta_1^* = R_{2|1}(R_{1|2}\beta_1) + (1 - R_{2|1})\beta_1^\dagger. \quad (8)$$

The remaining results given in the text are special cases of expressions 6 and 8: if the confounder is measured without error, $z_2 = x_2$, $R_{2|1} = 1$, and $\beta_1^* = R_{1|2}\beta_1$; if the variable of interest is measured without error, $z_1 = x_1$, $R_{1|2} = 1$, and

$$\begin{aligned} \beta_1^* &= \beta_1 + (1 - R_{2|1})(\sigma_{x_1,x_2}/\sigma_{x_1}^2)\beta_2 = \\ &R_{2|1}\beta_1 + (1 - R_{2|1})\beta_1^\dagger. \end{aligned} \quad (9)$$

For an alternative interpretation, note that $\sigma_{x_1,x_2}/\sigma_{x_1}^2$ is the slope $\gamma_{x_1|x_2}$ of the linear regression of x_1 on x_2 . In particular, expression 7 reduces if the variable of interest is measured without error to $\beta_1^\dagger = \beta_1 + \gamma_{x_1|x_2}\beta_2$, well-known as the effect of omitting a variable in linear regression (34).